Combined Shape Analysis of Human Poses and Motion Units for Action Segmentation and Recognition

Maxime Devanne, Hazem Wannous, Stefano Berretti, Pietro Pala, Mohamed Daoudi, Alberto del Bimbo

Combined Shape Analysis of Human Poses and Motion Units for Action Segmentation and Recognition

Maxime Devanne[1,2,3], Hazem Wannous[1], Pietro Pala[3], Stefano Berretti[3], Mohamed Daoudi[1,2], and Alberto Del Bimbo[3]
[1] University Lille 1 - CRIStAL (UMR CNRS 9189), France
[2] Institut Mines-Télécom/Télécom Lille, CRIStAL (UMR CNRS 9189) Lille, France
[3] Department of Information Engineering, University of Florence, Florence, Italy

***Abstract*— Recognizing human actions or analyzing human behaviors from 3D videos is an important problem currently investigated in many research domains. The high complexity of human motions and the variability of gesture combinations make this task challenging. Local (over time) analysis of a sequence is often necessary in order to have a more accurate and thorough understanding of what the human is doing. In this paper, we propose a method based on the combination of pose-based and segment-based approaches in order to segment an action sequence into motion units (MUs). We jointly analyze the shape of the human pose and the shape of its motion using a shape analysis framework that represents and compares shapes in a Riemannian manifold. On one hand, this allows us to detect periodic MUs and thus perform action segmentation. On another hand, we can remove repetitions of gestures in order to handle with failure cases for the task of action recognition. Experiments are performed on three representative datasets for the task of action segmentation and action recognition. Competitive results with state-of-the-art methods are obtained in both the tasks.**

I. Introduction

Analyzing and understanding human activities and behaviors is a problem that has been widely investigated in the past two decades. Indeed, this represents a task of interest for many promising applications in different domains, like surveillance, video games, physical rehabilitation, etc. Challenges appear when detecting humans and tracking their motion is required. Indeed, illumination changes or dynamic backgrounds can affect the human tracking and thus the understanding of his behavior. The emergence of 3D data allows capturing the human pose at each frame, thus reducing the challenges to human motion analysis. However, this task is still very difficult due to the temporal variability of behaviors, the complexity of human actions and the high number of motion combinations. In order to face these challenges, many works proposed to locally analyze the human behavior by decomposing and segmenting it into shorter and more understandable primitive motions.

Motion capture systems, like those from Vicon [14] are able of accurately capturing human pose, and track it along the time resulting in high resolution data, which include markers representing the human pose. Motion capture data have been widely used in industry, like in animation and video games. In addition, many datasets have been released providing such data for different human actions in different contexts, like the Carnegie Mellon University Motion Capture database [2].

More recently, new depth sensors have been released, like the Microsoft Kinect [10]. These new acquisition devices provide in addition to the classical RGB image, a depth image from which a 3D humanoid skeleton can be estimated thanks to the work of Shotton et al. [13]. Thus, such low-cost sensors offer a good alternative to capture human pose and human motion, which is more convenient for general public applications. This new type of data have stimulated the creation of human action datasets, like the MSR Action 3D dataset [8], and the MSRC-12 dataset [5], as well as the development of research works targeting human action recognition.

A. Related Work

In the literature, a lot of works have investigated the problem of segmenting a complex human motion sequence into distinct actions from both video data and 3D data. Most of the approaches based on 3D data use the motion capture data. Different kind of methods exist to address the problem of action segmentation. The first category includes methods based on the detection of changing points representing changing motions. For instance, in [1] a method based on probabilistic principal component analysis is proposed, which detects when the distribution of human poses change over time. In [4], they detect changing points by using the zero-velocity crossing points of the angular velocity of joints. Differently, temporal clustering methods try to group successive and similar poses into clusters resulting to a decomposition of the sequence into several segments belonging to one of the clusters. Such method allows combined action segmentation and recognition. For instance, in [19] a method is proposed to find the best segmentation of a sequence by minimizing the error across the segments belonging to the clusters. Finally, in [18] a fuzzy segmentation method is proposed to model gradual transitions between temporally adjacent actions instead of considering a fixed changing point.

Human action recognition from 3D data has attracted many research groups in the recent years, since the release of depth sensors providing skeleton data. Such skeleton data are either used lonely [9], [17] or in combination with raw

This work was not supported by any organization

data [11], [15] provided by depth sensors, like color or depth images. For instance, Zanfir et al. [17] propose a moving pose descriptor which captures both the geometric information about the human pose as well as its speed and its acceleration within a short time interval. In [9], they differently use skeleton data by considering pairwise relative positions between joints. Then, a sequence is represented with a constrained method based on dictionary learning and applied to the joints of the skeleton. The same features are used in [15], but in addition the Local Occupancy Patterns describing depth appearance around each joints are considered. Likewise, in [11] skeleton features based on joint angles and depth features computed with histogram of oriented gradients are combined together.

B. Overview of Our Approach

In this paper, we propose an approach based on shape analysis of both human pose and motion for the task of action segmentation and recognition. First, we analyze locally the shape of the human pose in order to detect its changes and decompose a motion sequence into different motion units (MUs) representing small motions performed by the subject. Then, we analyze the shape of such spatio-temporal MUs in order to detect possible repeated cycles. For the task of action segmentation, similar cycles are grouped to form longer segments representing actions. For the action recognition task, repetitions are ignored in order to improve the classification accuracy obtained in [3]. Indeed, we observed that a different number of repetitions of a gesture within an action sequence may affect the recognition. Representing all sequences with only one instance of the gesture allows us to deal with this issue. The shape analysis of human poses and MUs is performed on Riemannian shape space, considering 3D curves and on higher dimensional spatio-temporal trajectories, respectively.

The rest of the paper is organized as follows: Sect. II presents the Riemannian approach used to analyze and compare shapes of curves and trajectories; Sect. III and Sect. IV explain how we apply the shape analysis for human poses and MUs, respectively; In Sect. V, the approach is evaluated on data of two different types for the tasks of action segmentation and recognition. Finally, Sect. VI concludes the paper and discusses future research directions that we would like to investigate.

II. Shape Analysis of curves in $\mathbb{R}^n$

In this Section, we introduce the Shape Analysis framework used to analyze and compare shape of human poses, as well as shape of MUs. As explained later, this framework allows us to represent the shape of curves in $\mathbb{R}^n$ and provides an elastic metric representing similarities between shapes. Note that, in this work $n = 3$ for the case of human poses analysis, and $n = 3N_j$ for the case of MUs analysis, being N_j the number of joints of the skeleton.

Let $\beta : I \rightarrow \mathbb{R}^n$ representing a n-dimensional curve, normalized in the I = [0,1] interval. We mathematically represent the shape of β using the *Square-root Velocity Function* (SRVF) defined as:

$$q(s) \doteq \frac{\dot{\beta}(s)}{\sqrt{\|\dot{\beta}(s)\|}} \ . \tag{1}$$

The SRVF is a special function first introduced in [6] that captures the shape of β and thus allows shape analysis of curves. As shown in [6], with this representation, the elastic metric to compare shape of curves is reduced to the $\mathbb{L}^2$ metric. We define the set of all functions as:

$$\mathcal{C} = \{q : I \rightarrow \mathbb{R}^n | \ \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^n) \ . \tag{2}$$

By restricting the length of β to 1, the space $\mathcal{C}$ becomes an infinite dimensional unit-sphere representing the *pre-shape space* of all curves invariant to translation and uniform scaling, where each SRVF associated to a curve is viewed as an element of $\mathcal{C}$. Considering the $\mathbb{L}^2$ metric on its tangent space, $\mathcal{C}$ becomes a Riemannian manifold, as demonstrated in [6]. To compare two curves, we can compute a distance between their corresponding shape on $\mathcal{C}$, which is defined as the length of the geodesic connecting the two elements on $\mathcal{C}$. As $\mathcal{C}$ is a sphere, the geodesic length between two elements q_1 and q_2 is defined as:

$$\theta = d_{\mathcal{C}}(q_1, q_2) = cos^{-1}(\langle q_1, q_2 \rangle) \ . \tag{3}$$

The geodesic path between these two elements is defined as:

$$\alpha(\tau) = \frac{1}{\sin(\theta)} \left(\sin((1-\tau)\theta)q_1 + \sin(\theta\tau)q_2\right) \ . \tag{4}$$

Such a geodesic path represents the elastic deformation of the shape q_2 to correspond to q_1. In particular, $\tau \in [0, 1]$ in Eq. (4) allows us to parametrize the displacement along the geodesic path α: $\tau = 0$ and $\tau = 1$ correspond, respectively, to the extreme shapes q_1 and q_2; An intermediate value of τ corresponds to an intermediate deformed shape between q_1 and q_2. Thus, in addition to have a distance representing the similarity between two shapes, such a framework also provides geodesics connecting two shapes representing the optimal elastic deformation between them.

However, shape analysis usually requires invariance to different transformations, such as translation, scale, rotation and re-parametrization. By representing a curve using the SRVF, we deal with the translation and scaling variability. However, rotation and re-parametrization still remain. Indeed, if a curve is rotated or re-parameterized, its SRVF changes, but its shape remains unchanged. We define the rotation group $SO(3)$ and the re-parametrization group Γ, where elements $\gamma \in \Gamma$ are re-parametrization functions. Rotating a curve β with $O \in SO(3)$ and re-parametrizing it with $\gamma \in \Gamma$ results to a new curve $\beta' = O(\gamma \circ \beta)$ equivalent to β in term of shape. Another advantage of the SRVF is that the actions of the product group $SO(3) \times \Gamma$ on $\mathcal{C}$ is on isometries. Thus, the SRVF of $\beta' = O(\gamma \circ \beta)$ is given by $\sqrt{\dot{\gamma}(t)}O(q \circ \gamma)(t)$. We define the equivalence class of q as:

$$[q] = \{\sqrt{\dot{\gamma}(t)}O(q \circ \gamma)(t) | O \in SO(3), \ \gamma \in \Gamma\} \ , \tag{5}$$

where each element of $[q]$ is equivalent up to a rotation and a re-parametrization. The set of all equivalence classes is called the *shape space* denoted as $\mathcal{S}$. To compute the geodesic distance between $[q_1]$ and $[q_2]$ on $\mathcal{S}$, we first need to find the optimal rotation O^* and re-parametrization γ^* that best register the element q_2 with respect to q_1. In practice, Singular Value Decomposition is used to find optimal rotation, and Dynamic Programming is used to find optimal re-parametrization. Let q_2^* being the element associated with O^* and γ^*, then $d_{\mathcal{S}}([q_1],[q_2]) = d_{\mathcal{C}}(q_1, q_2^*)$.

In this way, a distance representing the similarity between the shape of curves in $\mathbb{R}^n$ is defined independently to their translation, scale, rotation and re-parametrization variabilities. Note that, as explained below, not all these invariances are required or some of them are handled differently in the context of shape analysis of human poses and MUs.

III. Pose-based approach

In this Section, we present our approach to analyze a sequence locally at the level of human poses. At this level, shape analysis of human pose allows segmenting a sequence into MUs using the provided elastic metric in shape space.

A. Pose Representation

To analyze the shape of the human pose, we propose to represent it as a curve and to use the shape analysis framework described in Sect. II. The 3D coordinates of each joint of the skeleton are used. By connecting the 3D joints, we can obtain a 3D curve representing the shape of the human body, as shown in Fig. 1. In order to keep natural information about the human shape represented by the limbs, we do not randomly connect the different joints, but keep a coherent structure linking together joints belonging to the same limb. Thus, a 3D curve representing the human pose connects successively the spine's joints, the arms's joints and the legs's joints. In this way, a human pose is represented by a 3D curve instead of a 3D skeleton. We can now perform shape analysis of curves using the shape analysis framework and the provided distance, as described in Sect. II, for $n = 3$. Note that, in this case, as we compare poses of a same sequence (same subject), the scale of skeletons is unchanged during a sequence. As a 3D curve connects joints in a predefined order, the parametrization of curves remains the same along a single sequence. Thus, we do not need to find optimal re-parametrization between two shapes before computing the distance. Figure 1 shows a geodesic path between two poses represented by their 3D curve.

B. Motion Segmentation

Using our pose-based approach, we can locally analyze the evolution of the human pose along an action sequence. Thus, in order to split automatically the continuous sequence into segments exhibiting basic motions, called Motion Units (MUs), we detect when the motion is changing. We assume that when a human is performing two successive motions, its speed becomes slower at the end of the first motion and at the beginning of the second one. This results to similar poses

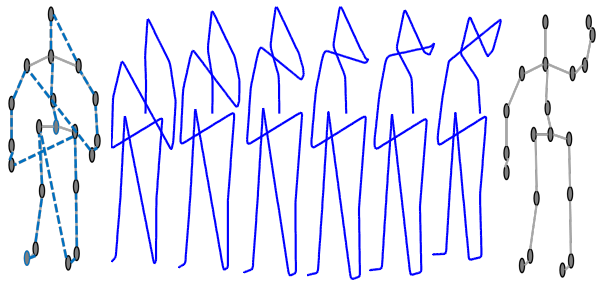

Fig. 1. A human pose is represented by a 3D curve. Geodesic distances can be computed between two poses in the *shape space*

in the time interval corresponding to the transition between the two motions. Our goal is to detect such transitions in an action sequence by analyzing human pose shape within a sliding window along the sequence. An important advantage of the shape analysis framework is that it allows the computation of statistics, like the mean and the standard deviation, on the manifold. Thus, for each window, we first compute the average pose corresponding to the Riemannian Center of Mass on the *shape space*, i.e., we use the distance $d_{\mathcal{S}}$ described in Sect. II. For the given shapes $q_1, \ldots, q_n$ corresponding to the poses $p_1, \ldots, p_n$ within a temporal window, their Riemannian Center of Mass is defined as:

$$\mu = \arg\min_{[q]} \sum_{i=1}^{n} d_{\mathcal{S}}([q],[q_i])^2 \ . \tag{6}$$

Once we have the mean shape, we compute the standard deviation between this mean shape and all the shapes in the window. The standard deviation is defined as:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_{\mathcal{S}}([\mu],[q_i])^2} \ . \tag{7}$$

Higher values of σ correspond to faster motion, while lower values correspond to slower motion, i.e., transition intervals. Figure 2 shows the evolution of σ along a sequence. We can easily detect different peaks corresponding to different MUs as shown in the skeleton sequence below the graph.

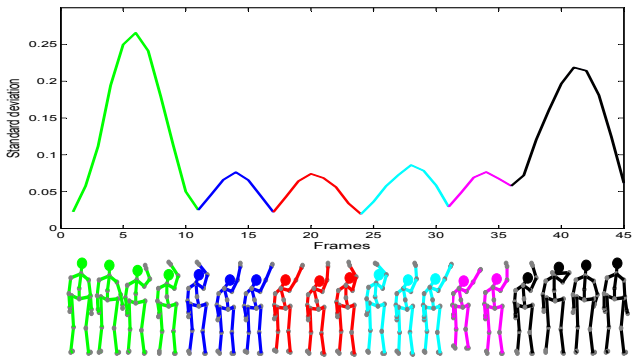

Fig. 2. Evolution of the standard deviation σ. Different peaks are detected and displayed with different colors. The corresponding poses are displayed under the plot

IV. Segment-based approach

With the approach defined in the previous section, we are able to decompose an action sequence into different segments or MUs that can be further analyzed. Based on this, in the following we describe our segment-based approach to perform action segmentation and action recognition.

A. Representation of MUs

Each MU represents the evolution of the human along a time interval. In order to capture both the geometric information about the human pose as well as the dynamic of the motion during the time interval, we represent the MU as a spatio-temporal trajectory of the human motion. A single skeleton of N_j joints is represented by a $3N_j$-dimensional vector by concatenating the three coordinates of each joint. Then, a feature matrix is built by concatenating the column vectors corresponding to each frame of the MU. This matrix represents the evolution of the human pose over time and can be viewed as a trajectory in a $3N_j$-dimensional space.

B. Detection of Periodic MUs

Looking at MUs as spatio-temporal trajectories, we can now use the shape space framework to analyze their shapes and compare them. In this case, higher dimensional space is considered ($n = 3N_j$). As a result, an elastic distance $d_{\mathcal{S}}$ can be computed, representing the similarity between MUs. We use this distance to detect repetitions of successive MUs. Note that, we first align MUs to a reference pose in order to analyze each MU independently to the orientation of the subject. As MUs are not necessarily repeated successively, but instead periodically, we search for different length of periodicity. Let MU_i be the i-th MU of a sequence and q_i its corresponding shape on the *shape space*. We define $P(\omega, i)$ the periodicity value of length ω for the i-th MU as:

$$P(\omega, i) = \frac{1}{\omega} \sum_{f=i-\omega}^{i} \phi(MU_f, MU_{f-\omega}) \,, \tag{8}$$

where:

$$\phi(MU_i, MU_j) = \begin{cases} 1 & \text{if } d_{\mathcal{S}}([q_i], [q_j]) < \text{threshold} \\ 0 & \text{otherwise} \end{cases} . \tag{9}$$

If $P(\omega, i) = 1$, a periodicity of length ω is detected at the i-th MU. We use this periodic detection for two different tasks: action recognition and periodic actions segmentation.

In the first case, a sequence contains one single action. In our previous work on action recognition [3], we represented an action sequence by a spatio-temporal trajectory. However, results demonstrated that this approach was unable to manage repetitions within a sequence. For instance, a *hammer* action can be performed more than once within a sequence, yielding trajectories with different shape. In this case, all training sequences are performed with only one instance of the action, while some test sequences are performed with several instances of the action. As we never trained a repeated action sequence, we are unable to recognize it. The method described above allows us to detect such repetitions. Thus, we finally represent the action by only one instance of the repetition. During the analysis of the sequence, if repetitions are detected, only the first instance is kept to represent the sequence. As a result, every sequence from the training or test set contains only one instance of the action. However, when repetitions are removed, we may loose the continuity of the action between the two remaining extreme parts of the sequence. In order to keep continuity, we use the two extreme poses (ending pose of the first part, and starting pose of the second part), represented in the *shape space*. Then, we estimate the deformation between these two poses using the geodesic path (Eq. 4). We discretize the path with a small number of steps representing the deformation between the two extreme poses. This process is illustrated in Fig. 3. Note that, the removed part of the sequence is a repetition of a previously observed MU. Thus, the ending poses should not differ a lot.

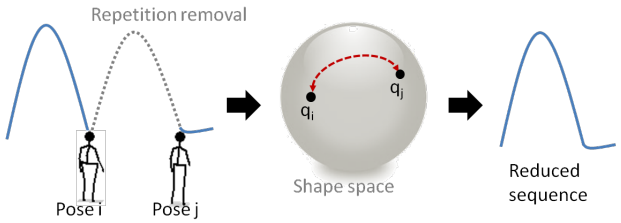

Fig. 3. Removal of repeated MUs keeping continuity of the action sequence. Deformation between the two extreme poses $Pose_i$ and $Pose_j$ is estimated using the geodesic path on the *shape space*

In the second case, a sequence contains successive periodic actions, such as *walking*, *running*, *boxing*, etc. The periodicity of the action is an important characteristic that allows us to perform segmentation. For instance, the action *walking* is a succession of *left step* and *right step*. Our method described in Sect. III allows the segmentation of the sequence in MUs corresponding to *left step* and *right step*. Once the segmentation is performed, we detect periodic MUs in order to group them in the same action cluster, e.g., *walking*. Thus, in this case, when repetitions are detected, we do not keep only the first instance, but group all instances in the same cluster. Detecting such periodic MUs along the whole sequence results in a segmentation of the sequence into different clusters, as illustrated in Fig. 4.

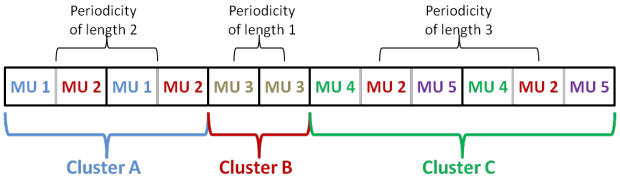

Fig. 4. Clustering of periodic MUs

V. Experimental results

We demonstrate the usefulness of the proposed approach for two different tasks: action segmentation and action recognition. First, we show how the detection of periodic MUs

is used for action segmentation. Second, we evaluate how the removal of repeated MUs improve our previous action recognition approach. The experiments are performed on three datasets, which provide data of two types: motion capture (mocap) data, in the CMU dataset; skeleton data captured with Microsoft Kinect in the MSR Action 3D and MSRC-12 datasets.

A. Action Segmentation

We evaluate the performance of our approach for the task of action segmentation using samples of the CMU dataset and compare it with the method proposed in [19] called HACA. Similarly to [19], we use 14 sequences performed by the subject #86. We evaluate the resulted segmentation in comparison with the ground truth by computing the confusion matrix between the segmentation obtained with our method and the ground truth. Then, we use the same metric used in [19] to compute the segmentation accuracy. Figure 5 shows the segmentation accuracy for the 14 sequences compared to [19]. It can be observed that we obtain competitive accuracies compared to HACA. Note that, a single sequence can include the same action several times at several time intervals, like *walking*. With our method, if a second instance of the same action happens, we view it as a new cluster. In order to handle this characteristic and be comparable with HACA, we assign the same label to similar clusters using the distance described in Sect. IV. Without this constraint, our approach segments a sequence in an online way parcouring only once the sequence with the sliding window method. In comparison, the offline method proposed in [19] needs a first initialization of the segmentation and then performs optimization in several iterations.

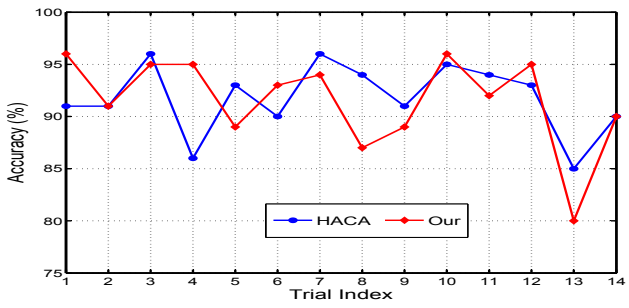

Fig. 5. CMU dataset: Segmentation accuracy for 14 sequences

Figure 6 shows the segmentation results of the fourth sequence obtained by HACA [19] (second row) and our method (third row), in comparison with ground truth segmentation (first row). Different colors correspond to different actions. The white bars within the same color represent the detected periodic movements. For instance, the first action in red (walking) is composed of five movements, each representing a walk cycle (one left step and one right step).

B. Action Recognition

1) MSR Action 3D Dataset: We demonstrate the usefulness of our approach to improve the action recognition of

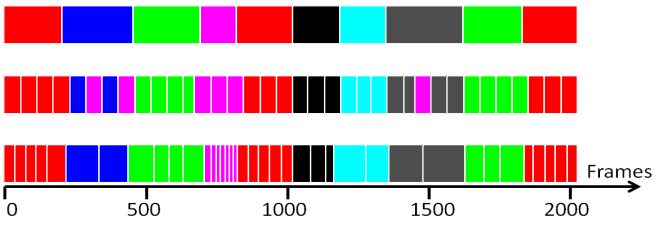

Fig. 6. Segmentation results obtained for a sequence. 1^{st} row corresponds to ground truth, 2^{nd} row to HACA [19] and 3^{rd} row to our approach

TABLE I
MSR ACTION 3D: COMPARISON OF THE PROPOSED APPROACH WITH THE MOST RELEVANT STATE-OF-THE-ART METHODS

Method	Accuracy (%)
Actionlet [15]	88.2
DCSF [16]	89.3
JAS & HOG^2 [11]	**94.8**
HON4D [12]	88.9
Moving Pose [17]	91.7
ScTPM [9]	93.8
Our previous [3]	92.1
Our	**94.3**

our previous work [3] evaluated on the MSR Action 3D dataset. We observed that if an action is repeated more than once within a sequence, it affects the shape of the corresponding trajectory, and thus the accuracy of the action recognition. We use the proposed segmentation approach to detect and remove such repetitions within a sequence. The overall accuracy is increased from 92.1% to 94.3%. However the experiments in [3] demonstrated that only one action class among 20 was mainly affected by this repetition variability (*hammer*). Table I shows that compared to state-of-the-art's method, such improvement allows us to obtain competitive accuracy.

2) MSRC-12 Dataset: We perform a third experiment on the MSRC-12 dataset including sequences of subjects performing 12 iconic and metaphoric gestures. In order to compare our method with [7], we only use the iconic gestures from this dataset. It results to 296 sequences of about 1000 frames length each, where a single gesture is performed several times along a sequence. The six classes are: *Duck*, *Goggles*, *Shoot*, *Throw*, *Change weapon* and *Kick*. 30 different persons perform each action several times resulting to about 50 sequences per class. Most of the cases, a gesture is repeated ten times within a sequence. However it may vary from 2 to 15. This point is very important to show how this variability can affect the recognition accuracy. In order to fairly compare our method with [7], we follow the same protocol. We employ a 5-fold leave-person-out-cross-validation, where each fold consists of 24 persons for training and 6 persons for test. Results are reported in Table II as average accuracies of each fold.

We can notice in Table II that we obtain lower accuracy than [7] only for one class (*Shoot*). The overall accuracy of our approach is 91.5%, which outperforms the one reported

TABLE II

MSRC-12: COMPARISON OF THE PROPOSED APPROACH WITH DFM [7]. ACCURACIES PER CLASS AS WELL AS MEAN ACCURACIES ARE REPORTED IN PERCENTAGE

Class	DFM [7]	Our previous [3]	Our
Duck	96.0	**100**	**100**
Goggles	88.0	82.0	**90.0**
Shoot	**85.7**	73.5	81.6
Throw	**90.0**	88.0	**90.0**
Change weapon	87.5	**89.6**	**89.6**
Kick	**98.0**	**98.0**	**98.0**
Mean	90.9	88.5	**91.5**

in [7] (90.9%). In addition, we can see that compared with our previous work without removing repetitions, the accuracy is increased. When we analyzed the failure cases, we noticed that the different number of repetitions within sequences affect the accuracy of our previous approach. This is for instance the case for similar actions, like *Goggles* and *Shoot*. If a test *Shoot* sequence includes a number of repetitions, which is not frequent in the the training *Shoot* sequences, but frequent in the training *Goggles* sequences, it may be assigned to a *Goggles* sequence and thus poorly classified. By representing each sequence as only one instance of the action, we are able to handle this issue.

To emphasize this point, we run a last experiment on a reduced version of the dataset. We only use sequences belonging to the classes *Goggles* and *Shoot*. In the training set, we first include *Goggles* sequences with exactly 10 repetitions of the gesture. Then, we include all *Shoot* sequences except those with exactly 10 repetitions of the gesture, which are included in the test set. We then evaluate the recognition accuracy of our previous method and with the improvement presented in this work. The recognition accuracy of the class *Shoot* is increased from 39.4% to 78.8%. This shows that our method allows us to improve the recognition accuracy when the number of repetitions of a single gesture can vary within a sequence.

VI. CONCLUSIONS

In this paper, we first presented an approach to decompose a sequence of body skeleton poses into a combination of motion units (MUs). By representing the human pose as a 3D curve, we analyze its shape on a Riemannian manifold and thus detect changes over time resulting to MUs. These MUs are then represented as spatio-temporal motion trajectories in order to analyze their shape similarly, and thus detect repetition of MUs. On one hand, we applied this method for the task of action segmentation by grouping successive similar motion units together. On another hand, we have shown how this method can improve the action recognition performance by representing an action sequence as only one instance of the action. Evaluation performed on three datasets providing two types of 3D data demonstrated that our method gives comparative results with respect to state-of-the-art work.

The obtained results motivated us to extend the idea of MUs to more complex cases, like activity recognition, where the complexity of the human motion is increased and where other aspects, like interaction with objects, make the recognition task more difficult. In particular, we plan to investigate a model capable of representing an activity sequence based on MUs performed by the subject as well as on the object held by the subject during the sequence.

REFERENCES

[1] J. Barbic, A. Safonova, J.-Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard. Segmenting motion capture data into distinct behaviors. In *Proc. Graphics Interface*, 2004.

[2] Carnegie Mellon University Motion Capture Database. `http://mocap.cs.cmu.edu`, 2012.

[3] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo. 3D human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Trans. on Cybernetics*, preprint, 2014.

[4] A. Fod, M. J. Mataric, and O. C. Jenkins. Automated derivation of primitives for movement classification. *Autonomous Robots*, 12(1):39–54, Jan 2002.

[5] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In J. A. Konstan, E. H. Chi, and K. Höök, editors, *CHI*, pages 1737–1746. ACM, 2012.

[6] S. H. Joshi, E. Klassen, A. Srivastava, and I. Jermyn. A novel representation for Riemannian analysis of elastic curves in R^n. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1–7, Minneapolis, MN, USA, June 2007.

[7] A. Lehrmann, P. Gehler, and S. Nowozin. Efficient nonlinear markov models for human motion. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1314–1321, Columbus, OH, USA, June 2014.

[8] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *Proc. Work. on Human Communicative Behavior Analysis*, pages 9–14, San Francisco, California, USA, June 2010.

[9] J. Luo, W. Wang, and H. Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1809–1816, 2013.

[10] Microsoft Kinect. `http://www.microsoft.com/en-us/kinectforwindows/`, 2013.

[11] E. Ohn-Bar and M. M. Trivedi. Joint angles similarities and HOG^2 for action recognition. In *Proc. CVPR Work. on Human Activity Understanding from 3D Data*, pages 465–470, Portland, Oregon, USA, June 2013.

[12] O. Oreifej and Z. Liu. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, pages 716–723, Portland, Oregon, USA, June 2013.

[13] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1–8, Colorado Springs, Colorado, USA, June 2011.

[14] Vicon Motion Systems. `http://www.vicon.com/`.

[15] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, Providence, Rhode Island, USA, June 2012.

[16] L. Xia and J. K. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proc. CVPR Work. on Human Activity Understanding from 3D Data*, pages 2834–2841, Portland, Oregon, USA, June 2013.

[17] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2752–2759. IEEE, 2013.

[18] H. Zhang and W. Zhou. Fuzzy segmentation and recognition of continuous human activities. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 6305–6312. IEEE, 2014.

[19] F. Zhou, F. De la Torre, and J. K. Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(3):582–596, Mar 2014.